GPTScore: Evaluate as You Desire

Year: 2023 | Citations: 382 | Authors: Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, Pengfei Liu

---

Abstract

Generative Artificial Intelligence (AI) has enabled the development of sophisticated models that are capable of producing high-caliber text, images, and other outputs through the utilization of large pre-trained models.Nevertheless, assessing the quality of the generation is an even more arduous task than the generation itself, and this issue has not been given adequate consideration recently.This paper proposes a novel evaluation framework, GPTScore, which utilizes the emergent abilities (e.g., in-context learning, zero-shot instruction) of generative pre-trained models to score generated texts. There are 19 pre-trained models explored in this paper, ranging in size from 80M (e.g., Flan-T5-small) to 175B (e.g., GPT3).Experimental results on four text generation tasks, 22 evaluation aspects, and corresponding 37 datasets demonstrate that this approach can effectively allow us to achieve what one desires to evaluate for texts simply by natural language instructions.This nature helps us overcome several long-standing challenges in text evaluation–how to achieve customized, multi-faceted evaluation without model training. We make our code publicly available.